The advantages of UK Biobank's open access strategy for health research

The advantages of UK Biobank's open access strategy for health research

Running title: Access to UK Biobank

Megan Conroy[1,2], Jonathan Sellors[1], Mark Effingham[1], Thomas J. Littlejohns[1,2], Chris Boultwood[1], Lorraine Gillions[1], Cathie L.M. Sudlow[1,3], Rory Collins[1,2], Naomi E. Allen[1,2]

[1] UK Biobank, Cheadle, Stockport, UK

[2] Nuffield Department of Population Health, University of Oxford, UK

[3] Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, UK

Abstract

Ready access to health research studies is becoming more important as researchers, and their funders, seek to maximise the opportunities for scientific innovation and health improvements. Large-scale population-based prospective studies are particularly useful for multidisciplinary research into the causes, treatment and prevention of many different diseases. UK Biobank has been established as an open-access resource for public health research, with the intention of making the data as widely available as possible in an equitable and transparent manner. Access to UK Biobank's unique breadth of phenotypic and genetic data has attracted researchers worldwide from across academia and industry. As a consequence, it has enabled scientists to perform world-leading collaborative research. Moreover, open access to an already deeply characterized cohort has encouraged both public and private sector investment in further enhancements to make UK Biobank an unparalleled resource for public health research and an exemplar for the development of open access approaches for other studies.

Keywords: epidemiology; public health; science

Introduction

Over the last few decades, several large-scale observational studies have been established to enable epidemiological research into the causes of the major diseases of middle and old age. Many of these studies express a commitment to open data sharing in order to facilitate research efforts, whilst ensuring appropriate commitment to participant confidentiality, consent and data protection regulations. This has become even more important in the era of genomics where meta-analyses of data from multiple (largely retrospective) studies are essential to achieve the numbers required to perform population-based genetic research [1, 2] and often requires collaboration with the team that set up the study. However, few epidemiological studies have been designed from the outset to be an open-access resource available to academic and commercial researchers alike from around the world, with no preferential access.

This article describes the access policy of UK Biobank, how it has developed over time in relation to both the use of data and biological samples, and how it has facilitated collaborative research whereby the results can be shared by all.

UK Biobank

UK Biobank is a large, prospective cohort study of 500,000 participants aged 40 to 69 years at the time of their baseline assessment visit during 2006-2010. The study was established to enable research into the lifestyle, environmental and genomic determinants of life-threatening and disabling diseases of middle and old age. A vast amount of data was collected at recruitment, including self-reported lifestyle and medical information (supplemented subsequently by antecedent information from health records), a wide range of physical measures (e.g. blood pressure, anthropometry, spirometry) and biological samples (blood, urine and saliva), of which further details are provided elsewhere [3]. All of the data can be viewed on UK Biobank's online Data Showcase, including summary statistics for each data-field available for research [4].

Since recruitment, UK Biobank has continued to be enhanced by converting the information contained in the biological samples, which are limited and depletable, into data that can be widely shared. This has included cohort-wide genotyping (with subsequent imputation to over 90 million variants) [5] and whole exome sequencing, making it one of the largest studies in the world with detailed data on genetics, lifestyle and health outcomes. A range of blood and urine biomarkers of interest for research into common conditions (such as cardiovascular disease, cancer and diabetes) are also available for all 500,000 participants [6]. UK Biobank continues to collect extensive data directly from participants. This includes a series of web-based questionnaires sent to all participants with an email address (n=330,000) about particular exposures (e.g. diet, occupation) and conditions (e.g. cognition, mental

health, pain), objective physical activity monitoring (100,000), and ongoing assessments of multi-modal imaging (target of 100,000) and cardiac monitoring (target of >20,000).

As UK Biobank is a prospective study, considerable efforts are spent in following the health of all participants through linkage to electronic health datasets, including death and cancer registries, and primary and secondary care records (Fig. 1). Several thousand incident cases of the most common conditions have already been identified, with many more cases expected to accrue over the next few years (Table 1). Efforts are underway to generate algorithmically-derived health outcomes in order to facilitate a wide range of research using standardised outcome variables [7].

Access to the resource

UK Biobank was set up on the basis of a clear intention from its two core funders (the Medical Research Council and Wellcome Trust) as a de-facto open-access resource, with the aim to make the data as widely available as possible, with an equitable and transparent access policy [8].

In order to apply for access to data from UK Biobank, each applicant must demonstrate that they are a *bona fide* researcher (i.e. they must register from, and be affiliated with, an approved research institute) and the application must involve health-related research that is in the public interest. All applicants are treated the same – whether academic, governmental, charitable or commercial, or whether from domestic or international organisations – and all applications are assessed according to the same consistent criteria.

All access applications are discussed and approved by the Access Sub-Committee (ASC) of the UK Biobank Board. Access to data is relatively permissive, and review by the ASC seeks only to ensure that the research is viable and meets the requirements. The ASC's main responsibility is making strategic access decisions, particularly regarding contentious matters and the use of biological samples. Ethics advice is provided to the ASC on an independent consultancy basis by Oxford University's Ethox group [9].

Lay summaries of each approved application are published on the website. A standard material transfer agreement (MTA) is signed prior to any data delivery, and governs how a researcher can use the data. All researchers must publish (or otherwise make publicly available) the findings of their research and return any derived data-fields, and the methods used to generate them, back to UK Biobank. These data are available to other registered researchers, thereby encouraging transparency and reproducibility in scientific methods.

UK Biobank is established as a charity with access charges (reviewed on a periodic basis) which are set at a level that covers the costs of managing the access

application process. In order to encourage use by potentially disadvantaged researchers, fees are subsidised for research groups from low and low-to-middle-income countries (assessed according to the current World Bank guidelines) and for student projects.

Evolution of UK Biobank's access approach

When the UK Biobank resource opened to researchers in April 2012, a relatively cautious approach to data access was taken. At that time, the application process consisted of two phases, a preliminary form (for early identification of projects not deemed compliant with UK Biobank's purposes) and a main form, each requiring separate payment and approval at various levels. This involved reviews from the scientific team to ensure the project was well-defined and health-related, the data analysts to ensure the selected data-fields were appropriate, UK Biobank's Principal Investigator (UKBPI) to make a final check, and the ASC to provide official assessment with approval or rejection (with a right of appeal).

Initially, researchers had to have a clear, well-defined research question with a focus on specific exposures and outcomes and justify their requests for data-fields. Datasets were restricted to only those data and participants that the researcher required (e.g., women only or specific case-control subsets). As the sheer size and depth of available data has increased, particularly following inclusion of the genotype data into the resource, the requirements have been relaxed to enable research that is broader in scope and often exploratory in nature (i.e. hypothesis-generating), with about one-third of research groups requesting the entire core dataset. As interest in the resource has grown over time (see Figs. 2.a and 2.b), UK Biobank further streamlined its approach when it launched a new access management system in February 2018 [10]. Interested researchers still have to register with UK Biobank in order to verify their research credentials, but the application comprises a single simplified form with easier selection of data-fields. In a further revision of the process, UK Biobank intends to provide the entire core dataset (excluding potentially identifying and particularly complex and/or large data) for each research project. It is anticipated this will substantially streamline the process further as it removes the requirement both for researchers to select each data-field and for UK Biobank to produce bespoke datasets.

Most data-only applications are fundamentally non-contentious (with 99% approved), so further streamlining efforts have led to delegation of approval to the scientific team, with the ability to escalate applications to the UKBPI and ASC if considered necessary. These changes have led to a shorter turnaround time for applications: the time from application submission to data release has reduced from 69 weeks in 2013 to 24 weeks by the end of 2018. It is intended that this will continue to be reduced following the provision of a default core dataset and the removal of an upfront payment stage, to be implemented in mid-2019.

Access to biological samples

Applications that request access to biological samples undergo much more stringent consideration, as the samples are a limited and depletable resource. The science behind the request is reviewed rigorously and external expert advice sought, where necessary [11]. When the resource was established, it was envisaged that access to the biological samples (blood, urine and saliva) for assays would be co-ordinated around case-control subsets "nested" within the whole cohort, as performed in virtually all previous prospective studies to date. However, it became apparent that this would not be the most efficient (or cost-effective) way of developing the resource for researchers to study the causes of many different health outcomes. This is because assays of samples in nested case-control comparisons based on different subsets of the participants preclude reliable comparisons across the full cohort. In contrast, generating assay data from biological samples for the entire cohort at one time facilitates good quality control by reducing measurement error and assay drift. This strategy also minimises sample depletion and is highly cost-effective since, in the long-term, the costs of conducting assays at one time for all of the participants are likely to be less than the costs of multiple retrievals. As such, requests for UK Biobank samples (which comprise 4% of all submitted applications; Fig. 2.c) are now only considered where they are undertaken on the whole (or a large subset) of the cohort, the assay data are applicable to a range of researchers, the assay method is well validated and uses minimal sample volume, and the laboratory can adhere to strict quality control measures [11].

Access to participants for third party studies

At recruitment to the study, participants consented to being re-contacted by UK Biobank. This includes communications to inform participants about the progress of the study (e.g., via an annual newsletter), and invitations to join third-party studies. As with samples, UK Biobank considers that re-contact of participants to be a depletable resource and is mindful not to over-burden participants with such invitations. Any application to use UK Biobank as a recruitment pool for third party studies (which comprise ~1% of all submitted applications; Fig. 2.c) is carefully reviewed by the ASC to ensure that there is sufficient scientific justification for such re-contact. As UK Biobank participants consented on the understanding that no results would be fed back to them following their assessment visits, care is taken to ensure that re-contact does not represent implicit feedback of information of which participants are not aware. As such, recruitment based on genotype or on phenotype that is not explicitly self-reported by the participant is highly restricted [12].

Who is using the data?

Since 2012, over 10,000 researchers have registered to use the resource, over 1,500 applications have been submitted and 1,000 projects are underway. The number of international researchers has steadily increased over time and now accounts for about three-quarters of all registrations and about two-thirds of all applications (Fig. 2.a and 2.b). Over 700 institutes worldwide have published using UK Biobank data. An independent analysis commissioned in 2018 highlighted that many non-UK institutes were using the resource with several major international groups – such as the Broad Institute/Harvard (USA), the University of Queensland (Australia), Erasmus University Medical Centre (Netherlands) and the Karolinska Institute (Sweden) – being particularly prolific users. True to the multidisciplinary nature of research, many research groups are collaborating with each other; for example, researchers from the Broad Institute/Harvard and the Universities of Oxford, Cambridge and Edinburgh frequently publish together, as do the Universities of Queensland and Edinburgh (Fig. 3.a).

The majority (>95%) of applications are for data-only (Fig. 2.c); true to the prospective nature of the resource, nearly all applications request death and cancer data, approximately three-quarters request the genomic data, two-thirds the hospital-inpatient data and one-third the imaging-derived phenotypes (i.e. variables generated from the raw imaging scans) (Fig. 2.d).

Growing interest from industry

The participant consent for UK Biobank is clear that access to the resource is available to commercial companies for use for health-related research on the same basis as academic researchers. Registered researchers from industry now account for 12% of all researchers as pharmaceutical and other commercial research groups realise the potential of the resource to accelerate drug discovery and develop machine-learning techniques for early detection of disease. Industry partners are also starting to enhance the resource further (for example, by supporting cohort-wide assays) in order to augment their own research aims, while at the same time benefiting the wider research community as the enhancements are shared with all researchers after a limited exclusivity period (now set at a fixed period of 9 months).

The first major industry investment was by Regeneron Pharmaceuticals to perform whole exome sequencing of the whole cohort. The first 50,000 samples have been sequenced in partnership with GlaxoSmithKline and these data are now available to all researchers. The remaining 450,000 samples are being exome sequenced by Regeneron in partnership with Abbvie, Alnylam, AstraZeneca, Biogen, Pfizer, Bristol-Myers Squibb and Takeda, and will be available to other researchers by the end of 2020. In addition, whole genome sequencing (WGS) is also underway on 50,000 participants, and it is anticipated that sequencing the remaining 450,000 participants

will be funded by a consortium of industry, government and charity funders, with data to become available to researchers over the next few years. In parallel, Nightingale Health, a biotech company from Finland, is measuring about 200 lipids and other circulating metabolites for all 500,000 participants. Government and charity funders have also provided funding for academic researchers to measure telomere length for all participants, and to collect data on heart arrhythmias via a dedicated heart monitor for 20,000 participants.

In addition, academic and industry collaborations are underway to process the raw scans collected as part of the ongoing imaging assessment of 100,000 participants in order to generate imaging-derived variables that can be used more readily by the wider research community. Because of the unprecedented scale of the imaging sub-study, this has necessitated the development of automated processing tools that can rapidly extract imaging-derived phenotypes. This includes phenotypes related to the structure and function of the brain (developed by The Wellcome Centre for Integrative Neuroimaging [13]), liver fat quantity and function (developed by Perspectum [14]), and detailed body composition measures (developed by several groups, including Advanced MR Analytics AB in conjunction with Pfizer [15], and Klarismo). These imaging-derived phenotypes are now being widely used by the wider research community to characterise intermediate disease outcomes and to investigate biological mechanisms of disease.

In this way, industry is effectively becoming a funder of UK Biobank, accelerating the rate at which the biological samples (e.g. through cohort-wide assays) and complex imaging data (e.g. raw magnetic resonance [MRI] scans) are converted into data that are potentially transformative in terms of the science they can support. Such large-scale investment is not feasible from most public sector sources, underscoring the effectiveness of UK Biobank's data sharing model.

Research output

The UK Biobank resource is generating an increasingly large and diverse research output related to identifying genetic and environmental risk factors for disease, with over 600 publications (Fig. 3.b) and over 10,000 citations (mostly in the last 2 years), as well as large numbers of conference abstracts, student projects, and methodological tools posted online.

The availability of genomic data on such large numbers is transforming genetic research, with Genome-Wide Association Studies (GWAS) now considered routine. Indeed, research groups have already made summary GWAS statistics for thousands of phenotypic traits publicly available [16-18]. This, in turn, is accelerating research into using genetic variants to assess causality of associations (e.g. using Mendelian Randomization approaches [19-21]) or for risk stratification purposes (e.g. using polygenic risk scores [22-25]). For the imaging research community, where

MRI data on this scale is unprecedented, both methodological and analytical advancements are underway to maximise the scientific utility of these data. For example, machine learning applications are being used to perform segmentation of MRI scans and to predict health outcomes [26].

Linkage to health data is allowing prospective analyses to be undertaken [27-29] and, as the cohort continues to mature, longitudinal research into the causes of a wide range of health outcomes will be possible. To date, cardiovascular, metabolic disease and cancer are the most common outcomes of research interest (Fig 3.c). However, this may well change as the numbers of incident cases of rarer conditions accrue over time. For example, 3,000 and 6,000 incident cases of osteoarthritis and hip fracture, respectively, will become available by 2022, enabling unprecedented research into their aetiology and progression (Table 1). In addition, the availability of primary care data in UK Biobank – which has hitherto not been available to UK cohort studies at a national level – will facilitate research into conditions (such as asthma, headaches, allergies, back pain, arthritis, diabetes, etc.) that are substantially under-ascertained when based only on hospital admission data. For example, the incorporation of primary care data in UK Biobank is anticipated to more than double the numbers of incident cases of chronic obstructive pulmonary disease (COPD) and dementia compared with hospital records and death data alone.

Data protection and de-identification

The processing and use of participant data are heavily regulated activities, particularly following the introduction of the General Data Protection Regulation (GDPR) in May 2018. This resulted in a specific communication to participants [30] setting out how the data that they had provided to UK Biobank were being used in accordance with the GDPR. Participant data provided to researchers are de-identified, so that potentially identifying information is either not released (e.g., name, NHS number) or is modified (i.e. home location grid co-ordinates are rounded to 1km; date of birth is restricted to month and year; certain brain images have facial features removed). UK Biobank is the only party that holds the necessary de-encryption keys to undertake re-identification, and different identifiers are used across different UK Biobank internal databases to protect against inappropriate re-identification (e.g., identifiable information is stored separately from phenotypic and genetic information; data collected during the imaging assessment have different identifiers to those of other data). Access to the keys that link the databases are highly restricted to designated staff to ensure the security of any identifiable data. Additionally, researchers agree when they sign the MTA prior to obtaining the data not to attempt to undertake re-identification of any participants for any purpose.

UK Biobank has a withdrawal process which allows a participant to withdraw from the resource at any time for any, or indeed no, reason. To date, since recruitment started, fewer than 800 participants have asked to be removed from future data collection (including linkage to electronic health records) and fewer than 200 have asked for their data and samples to no longer be available for research purposes.

Future direction: Dissemination of data

The growing volume of data associated with the increasing richness of the UK Biobank resource will inevitably drive changes in the way those data are disseminated. Hitherto, the approach to data distribution has involved researchers downloading data to their own local computing environment. This has already proved challenging in certain cases. For example, to ensure access for all researchers at exactly the same time, the genotyping data were initially made available in encrypted form and then de-encrypted simultaneously only when all researchers had had the opportunity to download them (so as not to disadvantage researchers with slower download capabilities).

The sheer volume of data associated with whole exome and whole genome sequencing of the entire cohort (currently estimated to be ~1PB and ~15PB, respectively) render unsustainable any approach based on distribution of data to researchers. UK Biobank is already starting to explore platform-based approaches, bringing researchers to the data rather than sending the data to researchers. By providing access to platforms which allow researchers to use the tools provided by the platform itself, or to run their own tools on the platform, the need to transfer data in bulk is avoided. Such a platform approach may also facilitate use of the UK Biobank resource by researchers at institutions that do not have a significant investment in local IT facilities, thus democratising further access to the data.

Conclusion

UK Biobank is being used by thousands of researchers worldwide for health-related research that is in the public interest. Its open-access strategy has enabled international scientists to produce excellent science and has led to external investment in enhancing the resource. As global interest in the resource grows, the data access process continues to be streamlined to enable researchers to obtain data quickly and easily. Open access of data to all researchers worldwide has encouraged both public and private investment, thereby enhancing this unique resource further.

Competing Interests:

All authors are current members of UK Biobank scientific team and/or executive management team. All authors have no conflicts of interest to declare.

Acknowledgements

UK Biobank is funded by the Medical Research Council, Wellcome Trust, Department of Health, Scottish Government, the Welsh Assembly Government, British Heart Foundation, and the Northwest Regional Development Agency. We would like to thank all the participants of UK Biobank for their vital contribution to the resource.

References

1. Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res* 2012; **40**: 3777-84.
2. Wellcome Trust Case Control Consortium (WTCCC). Available at: https://www.wtccc.org.uk/
3. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**: e1001779.
4. UK Biobank Showcase. Available at: http://biobank.ctsu.ox.ac.uk/crystal/
5. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; **562**: 203-09.
6. UK Biobank Biobank Biomarker Panel. Available at: http://www.ukbiobank.ac.uk/wp-content/uploads/2013/11/BCM023_ukb_biomarker_panel_website_v1.0-Aug-2015.pdf
7. Algorithmically-defined health outcomes in UK Biobank. Available at: http://biobank.ctsu.ox.ac.uk/crystal/docs/alg_outcome_main.pdf
8. UK Biobank protocol for a large-scale prospective epidemiological resource. Available at: http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf?phpMyAdmin=trmKQlYdjjnQIgJ%2CfAzikMhEnx6
9. Ethox Centre. Available at: https://www.ethox.ox.ac.uk/
10. UK Biobank's Access Management System. Available at: https://bbams.ndph.ox.ac.uk/ams/
11. UK Biobank's sample release policy and procedures. Available at: http://www.ukbiobank.ac.uk/wp-content/uploads/2017/12/Sample-release-policy-and-procedures.pdf
12. UK Biobank's re-contact procedures for third party researchers. Available at: http://www.ukbiobank.ac.uk/wp-content/uploads/2018/05/ukb-recontactprocs-14.3.2018-item-5b-2.pdf

13. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* 2016; **19**: 1523-36.
14. Hutton C, Gyngell ML, Milanesi M, Bagur A, Brady M. Validation of a standardized MRI method for liver fat and T2* quantification. *PLoS One* 2018; **13**: e0204175.
15. West J, Dahlqvist Leinhard O, Romu T, Collins R, Garratt S, Bell JD, et al. Feasibility of MR-Based body composition analysis in large-scale population studies. *PLoS One* 2016; **11**: e0163332.
16. Neale's Lab. Available at: http://www.nealelab.is/uk-biobank/
17. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nat Genetics* 2018; **50**: 1593-99.
18. McInnes G, Tanigawa Y, DeBoever C, Lavertu A, Olivieri JE, Aguirre M, et al. Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* 2018; bty999.
19. Gan W, Clarke RJ, Mahajan A, Kulohoma B, Kitajima H, Robertson NR, et al. Bone mineral density and risk of type 2 diabetes and coronary heart disease: A Mendelian randomization study. *Wellcome Open Res* 2017; **2**: 68.
20. He Y, Timofeeva M, Farrington SM, Vaughan-Shaw P, Svinti V, Walker M, et al. Exploring causality in the association between circulating 25-hydroxyvitamin D and colorectal cancer risk: a large Mendelian randomisation study. *BMC Med* 2018; **16**: 142.
21. Liu J, Rutten-Jacobs L, Liu M, Markus HS, Traylor M. Causal Impact of Type 2 Diabetes Mellitus on Cerebral Small Vessel Disease: A Mendelian Randomization Analysis. *Stroke* 2018; **49**: 1325-31
22. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol* 2018; **72**: 1883-93.
23. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Hoan Choi S, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genetics* 2018; **50**: 1219-24.
24. Rutten-Jacobs LC, Larsson SC, Malik R, Rannikmae K, Sudlow CL, Dichgans M, et al. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306,473 UK Biobank participants. *BMJ* 2018; **363**: k4168.
25. Smith T, Gunter MJ, Tzoulaki I, Muller DC. The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. *Br J Cancer* 2018; **119**: 1036-39.
26. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018; **2**: 158-64.
27. Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. *Lancet* 2015; **6736**: 1-8.

28. Millett ERC, Peters SAE, Woodward M. Sex differences in risk factors for myocardial infarction: cohort study of UK Biobank participants. *BMJ* 2018; **363**: k4247.
29. Pilling LC, Tamosauskaite J, Jones G, Wood AR, Jones L, Kuo CL, et al. Common conditions associated with hereditary haemochromatosis genetic variants: cohort study in UK Biobank. *BMJ* 2019; **364**: k5222.
30. UK Biobank GDPR. Available at: https://www.ukbiobank.ac.uk/gdpr/
31. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol* 2009, **38**: 263-73.

Figure legends

Fig. 1: Timeline of data collection and availability for UK Biobank participants by mid-2019. Pie chart indicates the proportion of the cohort that each data item is available for.

[a] Data on exome sequencing data (for 50,000 participants) and serological markers of infectious agents (for 10,000 participants) were made available in March 2019, with the intention to assay all 500,000 participants over the next few years.

Fig. 2: Access metrics. (a) Number of international and UK researchers by year (b) Number of applications by year and country (c) Proportion of different types of submitted applications (d) Proportion of applications different types of data

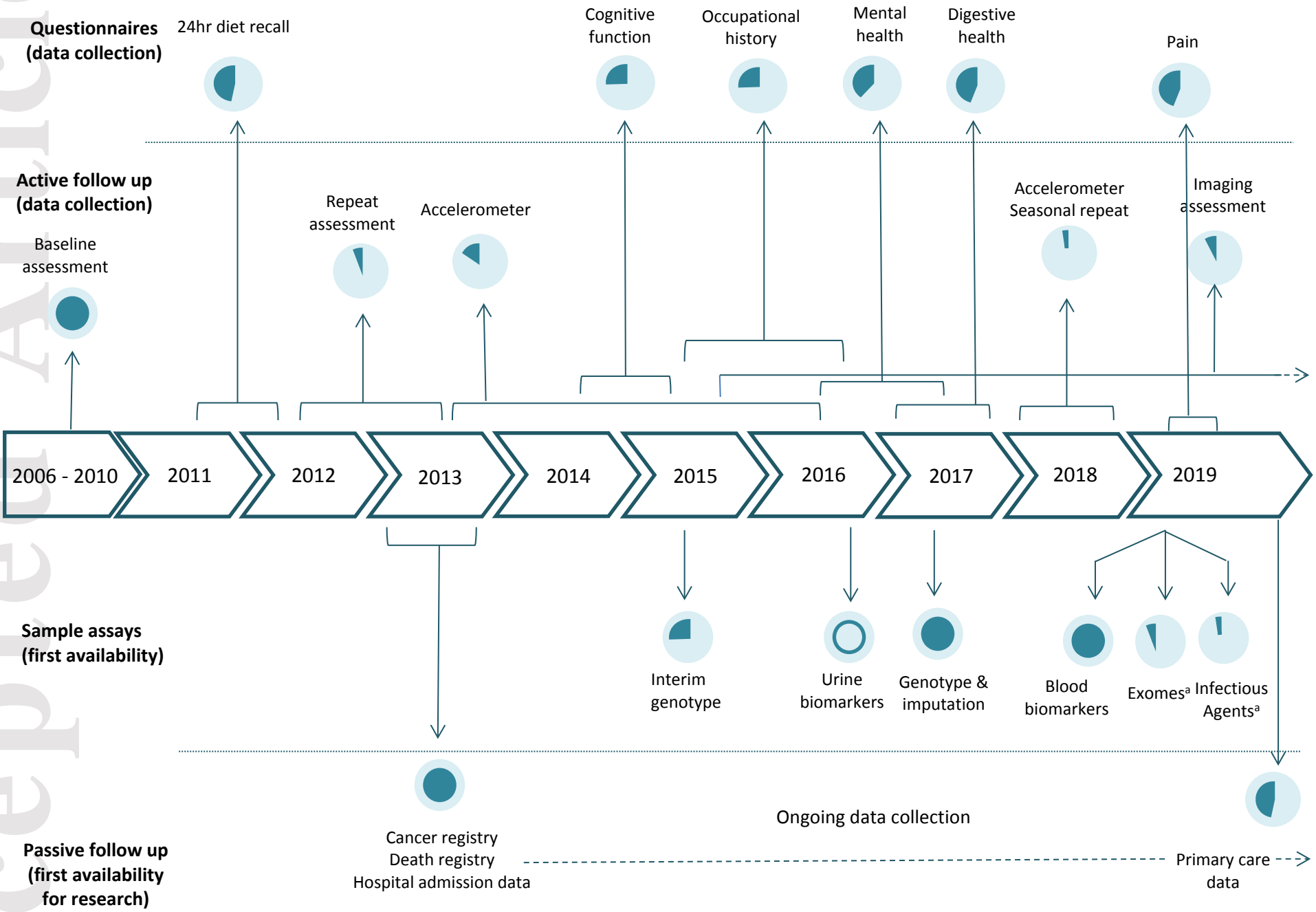
Fig. 3: Research metrics. (a) Collaborations between the top 12 institutes (b) Number of publications by year (c) Areas of research output.

[a]Graph generated by Digital Science & Research Solutions Ltd

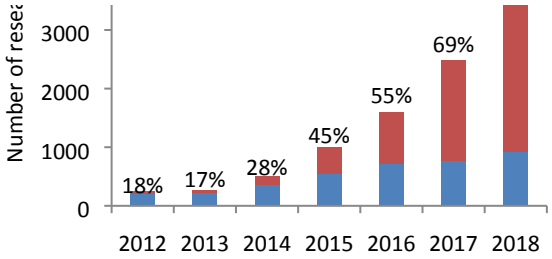
Table 1. Observed and expected numbers of selected health outcomes in UK Biobank over time[a]

Condition	Incident cases observed by 2016[a]	Incident cases predicted by 2026[b]
Dementia	4,300	43,400
Stroke	7,100	28,400
Myocardial Infarction	8,000	22,000
Chronic Obstructive Pulmonary Disease	17,600	55,000
Parkinson's Disease	2,000	9,700
Breast cancer	7,000	18,000
Prostate cancer	6,700	26,800
Colorectal cancer	4,000	16,000

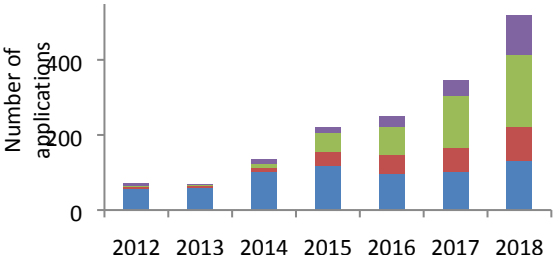
[a] Based on linkage to hospital in-patient records, death certificates, cancer registries and primary care (the latter extrapolated to the full cohort) up until 01 Jan 2016. [b] Predicted numbers of cases were derived by applying ratios from a previous modelling exercise conducted for UK Biobank [31], which was based on UK age-specific disease incidence rates, adjusted to take account of the numbers of disease cases observed so far in UK Biobank participants (who have lower rates of most diseases compared with similar aged people in the general UK population) in linked healthcare data from primary and secondary care sources.



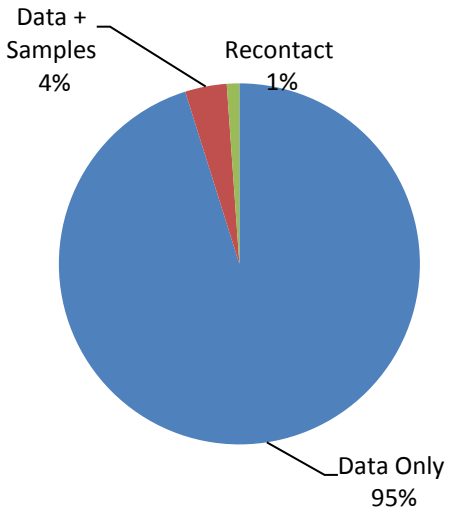
2a. Number of UK and international registered researchers by year



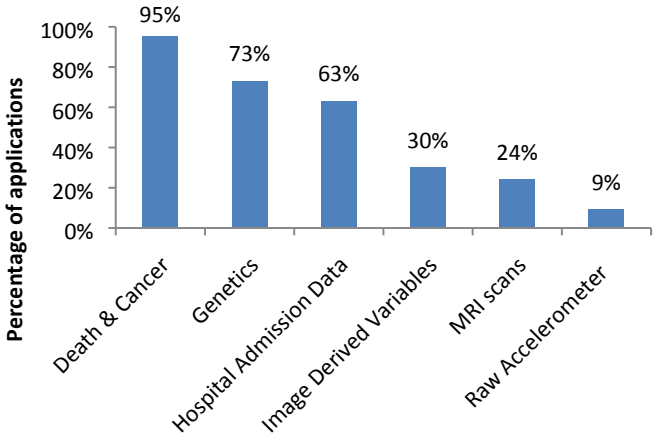
2b. Number of submitted applications by country and year



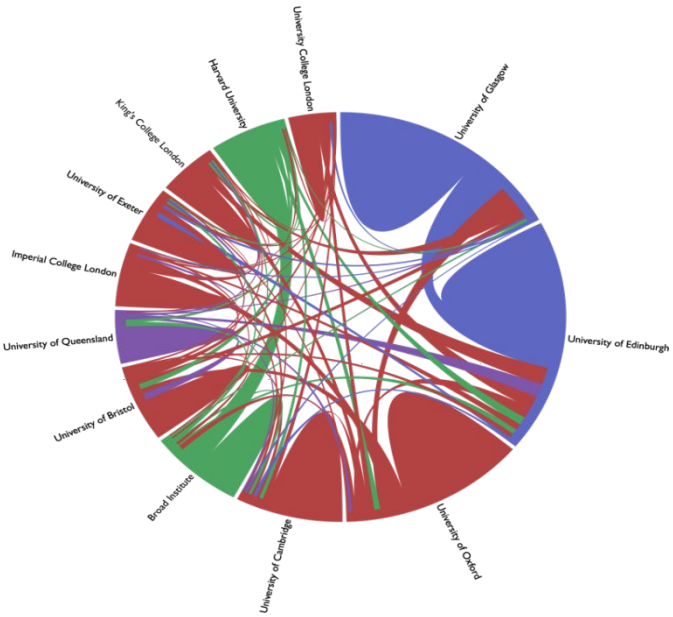
2c. Proportion of different types of submitted applications



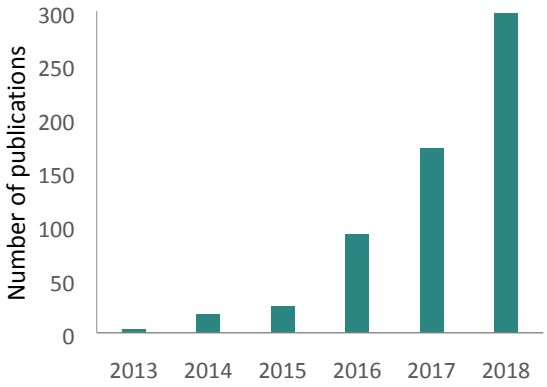
2d. Proportion of applications requesting different types of data



3a. Collaborations between the top 12 institutions using UK Biobank data[a]



3b. Number of publications by year



3c. Area of research output